

# Identification of Literary Movements Using Complex Networks to Represent Texts

Diego Raphael Amancio<sup>1</sup>, Osvaldo N. Oliveira Jr.<sup>1</sup>, Luciano da Fontoura Costa<sup>1</sup>

<sup>1</sup> Institute of Physics of São Carlos  
University of São Paulo, P. O. Box 369, Postal Code 13560-970  
São Carlos, São Paulo, Brazil

E-mail: [diego.amancio@usp.br](mailto:diego.amancio@usp.br), [diegoraphael@gmail.com](mailto:diegoraphael@gmail.com)

**Abstract.** The use of statistical methods to analyze large databases of text has been useful to unveil patterns of human behavior and establish historical links between cultures and languages. In this study, we identify literary movements by treating books published from 1590 to 1922 as complex networks, whose metrics were analyzed with multivariate techniques to generate six clusters of books. The latter correspond to time periods coinciding with relevant literary movements over the last 5 centuries. The most important factor contributing to the distinction between different literary styles was the average shortest path length (particularly, the asymmetry of the distribution). Furthermore, over time there has been a trend toward larger average shortest path lengths, which is correlated with increased syntactic complexity, and a more uniform use of the words reflected in a smaller power-law coefficient for the distribution of word frequency. Changes in literary style were also found to be driven by opposition to earlier writing styles, as revealed by the analysis performed with geometrical concepts. The approaches adopted here are generic and may be extended to analyze a number of features of languages and cultures.

<i>CONTENTS</i>	2
-----------------	---

## Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Modeling Texts as Complex Networks</b>	<b>3</b>
2.1	Pre-Processing . . . . .	3
2.2	Complex Networks Measurements . . . . .	4
<b>3</b>	<b>Database</b>	<b>7</b>
<b>4</b>	<b>Results and Discussion</b>	<b>7</b>
<b>5</b>	<b>Conclusion and further work</b>	<b>13</b>

PACS numbers: 89.75.Hc,89.20.Ff,02.50.Sk

## 1. Introduction

Many findings related to language and culture issues have been made with the use of statistical methods to treat large amounts of texts [1, 2, 3, 4]. Recent examples are the analysis of millions of books [1] and the study of twitter messages, where the global variation of mood could be observed through textual analysis of tweets [2]. In several of such examples knowledge is inferred from the analysis of semantic contents in the texts. There are also other methods to analyze text, including cases where text is represented as a graph (or network) [5]. Of particular relevance was the finding that networks formed from texts are scale free [6], whose topology could be analyzed leading to various contributions. For instance, the scale-free structure (which is analogous to the Zipf's Law frequency distribution [7]) of text networks emerged as a consequence of an optimization process for both hearer and speaker, so that the effort to transmit and obtain a message was minimized [8]. In addition to allowing for cultural features to be identified and explored, automatic analysis may be useful for real-world applications, such as automatic text summarization [9], machine translation [10, 11], authorship attribution [12], information retrieval [13] and search engines [14].

In this study we used topological metrics of complex networks representing text from 77 books dating from 1590 to 1922 in an attempt to verify changes in writing style. With multivariate statistical analysis of the metrics obtained, we were able to identify periods that correspond to major literary movements. Furthermore, we established which network characteristics were responsible for the changes in writing style.

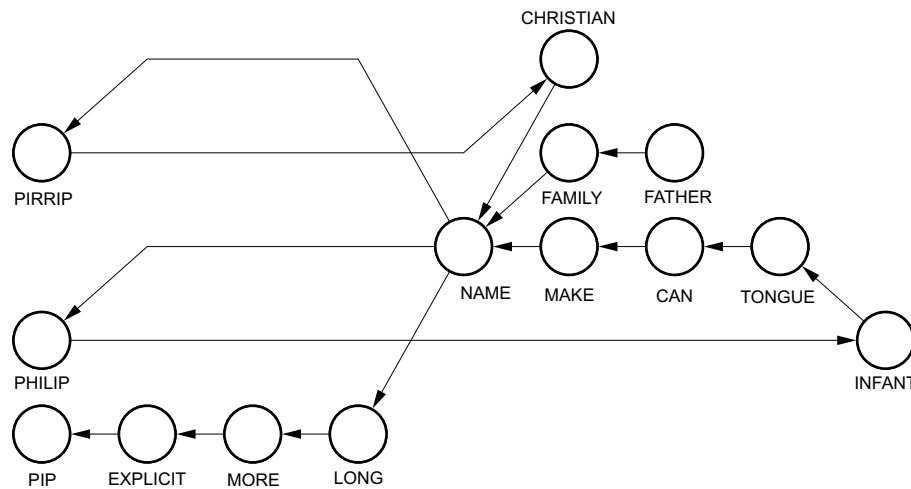
## 2. Modeling Texts as Complex Networks

### 2.1. Pre-Processing

The modeling process starts by removing punctuation and words that convey little semantic content (see the Supplementary Information (SI)-Sec.1), such as articles and prepositions. Then, the remaining words are transformed into their canonical form, i.e. nouns and verbs are converted into the singular and infinitive forms, respectively. This step is performed using the MXPOST part-of-speech tagger [15], which assists the resolution of ambiguities. The transformation to the canonical form (lemmatization) is done to cluster words referring to the same concept into a single node of the network despite the differences in flexion. At last, adjacent words in the written text are connected in the network according to the natural reading order (the left word is the source node and the right word is the target node). The modeling is demonstrated in Table 1 for the pre-processing steps, while Fig. 1 illustrates the network obtained from a small extract of the book *Great Expectations*, by Charles Dickens.

**Table 1.** Illustration of the pre-processing (removal of stopwords and punctuation marks) and lemmatization of the extract “My father’s family name being Pirrip, and my Christian name Philip, my infant tongue could make of both names nothing longer or more explicit than Pip.” obtained from the book *Great Expectations*, by Charles Dickens.

Original	Without stopwords	After lemmatization
My father’s family name	father family name	father family name
Pirrip, and my ,	Pirrip	Pirrip
Christian name Philip	Christian name Philip	Christian name Philip
my infant tongue	infant tongue	infant tongue
could make of both	could make both	can make both
names nothing longer	names longer	name long
or more explicit than Pip	more explicit Pip	more explicit Pip



**Figure 1.** Network obtained from the extract “My father’s family name being Pirrip, and my Christian name Philip, my infant tongue could make of both names nothing longer or more explicit than Pip.” of the book *Great Expectations*, by Charles Dickens.

## 2.2. Complex Networks Measurements

Several metrics extracted from the networks were used to quantify the style of the books. From each local measurement (i.e., which refers to a node) we derived some quantities describing the distribution of the networks in order to quantify the style of whole books. The measurements and their corresponding distribution descriptors were chosen because they have been useful to quantify the style of texts in previous studies [12]. The simplest measurement refers to the number  $N$  of nodes in the network, which corresponds to the size of the vocabulary used to write the piece of text analyzed. The distribution of word

frequency was characterized using the coefficient  $\gamma$  of the frequency distribution  $p_k$ :

$$p_k \sim ck^{-\gamma}, \quad (1)$$

where  $c$  is a normalization constant (see Fig. 2(a) for an example of the frequency distribution  $p_k$  of a specific book). We did not verify explicitly whether the degree obeys a power-law distribution because  $k$  is proportional to the frequency of words. Since the word frequency follows the Zipf's Law [16, 17], the degree is guaranteed to obey a power-law distribution<sup>‡</sup>. To compute  $\gamma$ , we employed a technique based on the accumulated distribution  $p_k$  (see Fig. 2(b)) described in Ref. [18]. We also used the frequency of words (or equivalently the degree  $k$  of the nodes) to calculate the assortativity  $\Gamma$  [19, 20, 21] (or degree-degree correlation) of the network as:

$$\Gamma = \frac{\frac{1}{M} \sum_{j>i} k_i k_j a_{ij} - \left[ \frac{1}{M} \sum_{j>i} \frac{1}{2} (k_i + k_j) a_{ij} \right]^2}{\frac{1}{M} \sum_{j>i} \frac{1}{2} (k_i^2 + k_j^2) a_{ij} - \left[ \frac{1}{M} \sum_{j>i} \frac{1}{2} (k_i + k_j) a_{ij} \right]^2} \quad (2)$$

where  $M = 21,900$ <sup>§</sup> is the number of edges of the network and  $a_{ij} = 1$  if nodes  $i$  and  $j$  are connected and  $a_{ij} = 0$  otherwise. If positive values are obtained for  $\Gamma$ , then highly connected nodes are usually connected to other highly connected nodes, indicating that there may exist regions where nodes are highly interconnected [19]. Conversely, if  $\Gamma$  is negative then highly connected nodes are commonly connected to little connected nodes.

In addition to measurements based on the number of nodes of the network and on the degree, the distance between concepts was employed to characterize the structure of the books. This measurement, widely known in the theory of networks as average shortest path length  $l$  [22], is calculated from the distance  $d_{ij}$ , which represents the minimum cost (minimum number of edges) required to reach node  $j$ , starting from node  $i$ . After computing all pairs of values  $d_{ij}$ , the average shortest path length  $l_i$  of each node  $i$  is:

$$l_i = \frac{1}{N-1} \sum_{j \neq i} d_{ij}. \quad (3)$$

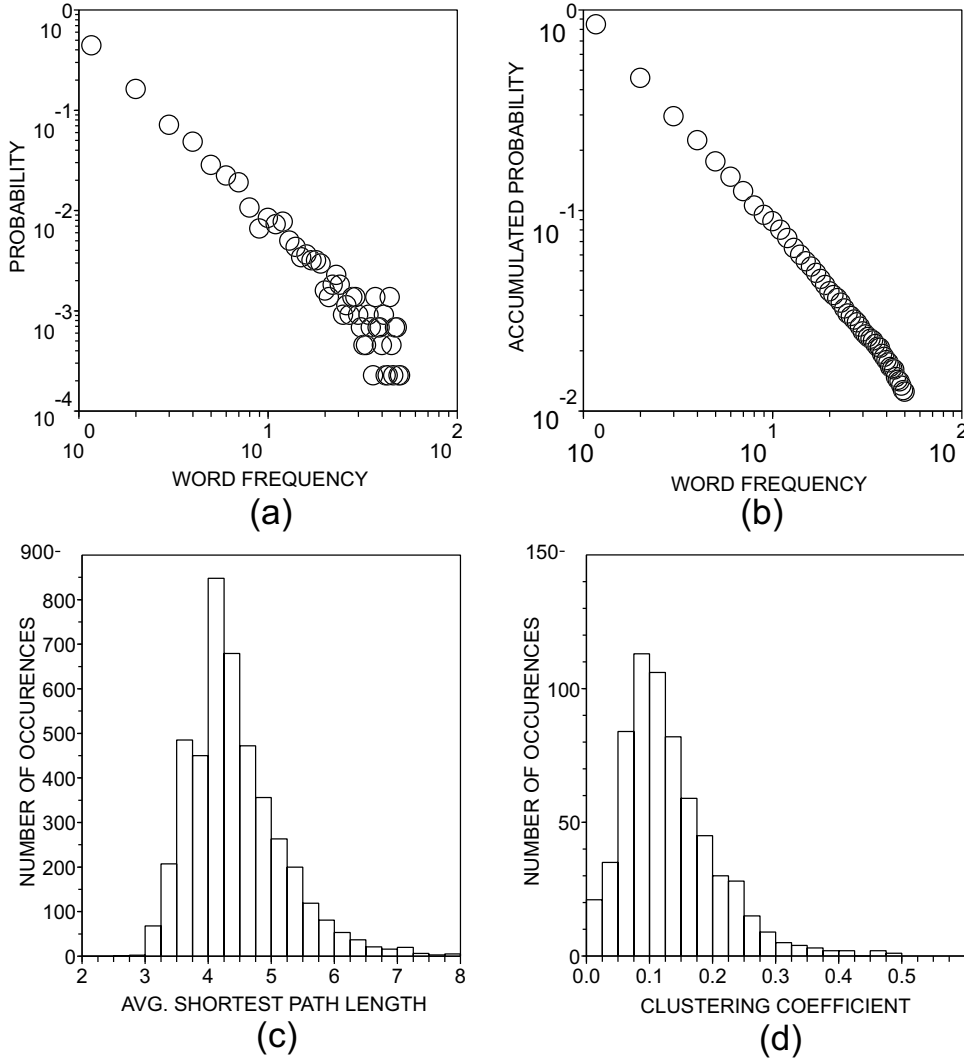
Since  $l_i$  is defined for each node individually, the network is characterized by a distribution of  $l_i$  (see the distribution of  $l_i$  for a specific book in Fig. 2(c)). The distribution was characterized quantitatively by computing the average  $\langle l \rangle$  and standard deviation  $\Delta l$ . Additionally, we computed the weighted average  $(1/\sum k_i) \sum k_i l_i \equiv \langle l_w \rangle$ , so that greater importance was given to the most frequent words in the text. The third moment  $\varsigma(l)$

$$\varsigma(l) = \frac{1}{N} \sum_{i=1}^N \left( \frac{l_i - \bar{l}}{\Delta l} \right)^3 = \frac{1}{N(\Delta l)^3} \left( \sum_{i=1}^N l_i^3 - 3\bar{l} \sum_{i=1}^N l_i^2 + 2N\bar{l}^3 \right) \quad (4)$$

was also computed.

<sup>‡</sup> The power-law distribution was verified for all texts of the database.

<sup>§</sup> To avoid effects from the size of the books, for obtaining the complex network we used only the first  $M + 1$  words of each book.



**Figure 2.** Example of distributions of measurements for the book *Great Expectations*, by Charles Dickens. The measurements used were: (a) simple word frequency; (b) accumulated word frequency; (c) average shortest path length; and (d) clustering coefficient. The adjusted R-square found in (a) was equal to 0.9348, which confirms that the frequency distribution is very similar to a power law distribution.

The last metric was the clustering coefficient ( $C$ ) [22], which quantifies the density of connections between the neighbors of a node  $i$  according to:

$$C_i = \frac{3 \sum_{k>j>i} a_{ij}a_{ik}a_{jk}}{\sum_{k>j>i} a_{ij}a_{ik} + a_{ji}a_{jk} + a_{ki}a_{kj}}. \quad (5)$$

The clustering coefficient in equation 5 represents the fraction of the number of triangles among all possible connected sets of three nodes, and therefore  $0 \leq C_i \leq 1$ . Similarly to the average shortest path length, it is also necessary to quantitatively characterize the distribution of the measurement (see an example of distribution of  $C$  in Fig. 2(d)). We therefore computed the average  $\langle C \rangle$ , the standard deviation  $\Delta C$ , the weighted average

$(1/\sum k_i) \sum k_i C_i \equiv \langle C_w \rangle$  and the third moment  $\varsigma(C)$  to characterize the distribution.

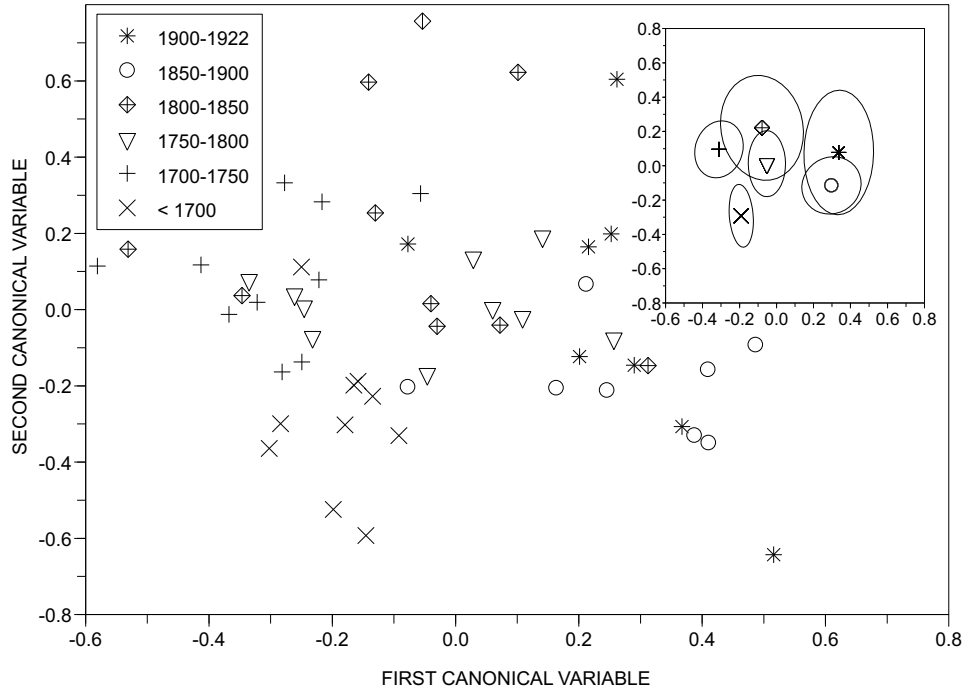
### 3. Database

The database comprises 77 books available online at the Gutenberg project repository [23], whose publication date ranged from 1590 to 1922. Tables S1-S3 in (SI)-Sec.2 give the details of the books. The texts were represented with complex networks [8, 9, 10, 11, 24, 25, 26, 27, 28, 29, 30], in which the edges are defined on the basis of co-occurrence of words (see Sec. 2). The latter procedure has been proven suitable to quantify both the style and structure of texts (see e.g. Refs. [11, 26, 29]). The details of the procedures adopted to model texts as complex networks and a description of the measurements employed to characterize the networks are given in Section 2.

### 4. Results and Discussion

The evolution of literary styles was quantified considering the 11 measurements from complex networks described in Sec. 2.2 for the books from the Project Gutenberg [23]. The main measurements were the shortest path length ( $l$ ), the clustering coefficient ( $C$ ), the assortativity ( $\Gamma$ ), the power law coefficient of the degree distribution ( $\gamma$ ) and the size of the vocabulary ( $N$ ). An initial, arbitrary division of the books in 6 intervals of 50 years, according to their publication date, led to the clusters shown in the Canonical Variate Analysis (CVA, see details in (SI)-Sec.3) plot in Fig. 3. The distinction was relatively poor, especially considering the standard variation ellipses [31] in the inset of the figure. Good separation was only possible when distant periods in time were compared, as their ellipses did not overlap. This difficulty in distinguishing literary movements should perhaps be expected as there is no reason for sharp transitions to occur only because half century marks were reached. We also verified the distinguishability of clusters with the Principal Component Analysis (PCA, see (SI)-Sec.3), but the distinction was also poor.

In order to verify whether books from distinct publication dates could be distinguished at all, we adopted a systematic procedure for the partition of the dataset using an optimization approach. This was performed by assessing the quality of the clustering under the condition that books with consecutive publication dates should belong either to the same cluster or lie in the boundaries of consecutive clusters. More specifically, we varied the delimiters and number of clusters in the database and quantified the quality of the clustering using 2 indices, viz. the simplified silhouette (SWC) and the Dunn index (DN) (see (SI)-Sec.4). Good distinction of writing styles was obtained for 3, 4, 5, 6 and 7 clusters (see Figure S1 of the SI), according to the two indices (SWC and DN). The best partition, which was found to be statistically significant (see Figure 4), was obtained with SWC and CVA projection, leading to the 6 clusters in Fig. 5, where there is almost no overlap among clusters, as shown in the inset. Most significantly, the 6 time periods inferred from this analysis coincide with



**Figure 3.** Scatter plot (CVA projection) representing the style of each book using 6 literary styles. Each style is represented by a set of 10 books. The inset displays the dispersion of the literary styles.

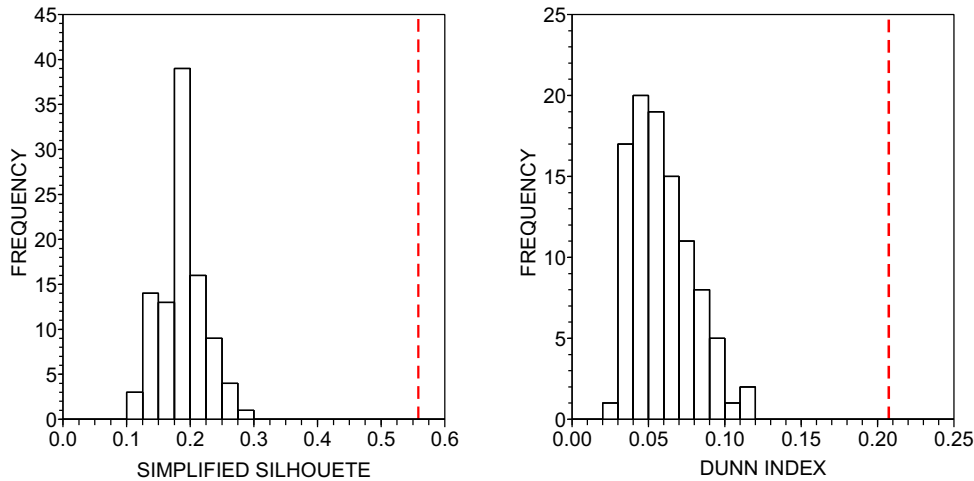
well-established literary movements listed in Table 2.

**Table 2.** Relationship between the best clustering of writing styles the traditional classification of literary movements.

Cluster Boundary	Literary Boundary	Literary Movement	Reference
1590 - 1653	1558 - 1603	Elizabethan era	[33]
1664 - 1761	1660 - 1798	Neoclassicism/Enlightenment	[34, 35, 36]
1767 - 1793	1660 - 1798	Neoclassicism/Enlightenment	[34, 37]
1794 - 1818	1764 - 1820	Gothic fiction	[34, 37]
1826 - 1906	1830 - 1900	Realism	[34]
1826 - 1906	1865 - 1900	Naturalism	[34, 38]
1906 - 1922	1890 - 1940	Modernism	[34, 39]

Other important features are inferred from Fig. 5. First, clusters for subsequent time periods are normally placed next to each other, indicating smooth changes in writing style over time. The same conclusion can be inferred from the analysis of the hierarchical clustering in Fig. 6 with the Wards [32] distance. The exception to this trend was the major change from the 1794 – 1818 → 1826 – 1906 period, which may



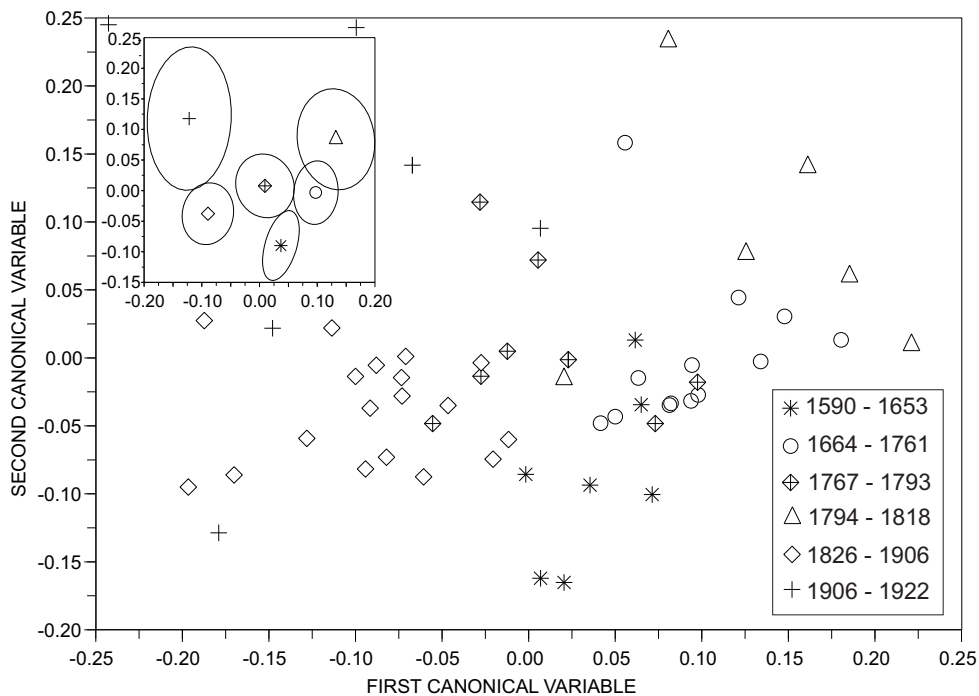


**Figure 4.** Significance test performed for (a) the simplified silhouette and for (b) the Dunn Index. The histograms represent the values of the cluster quality indices considering a random distribution of points and the dotted lines represent the clustering quality indices obtained for the clustering illustrated in Figure 5. Because the silhouette for the random case  $SWC_{rand} = 0.187 \pm 0.036$  is smaller than the silhouette  $SWC = 0.558$  for the clustering of Figure 5, the clustering inferred is significant. The same applies for the Dunn index because  $DN_{rand} = 0.059 < DN = 0.207$ .

be the consequence of a drastic change in style triggered by the French Revolution (1789). As for the variance among clusters, the lowest and highest values applied to the 1590 – 1653 and 1906 – 1922 periods, respectively. These results are intuitive as little change in style could be expected in older periods, while in the recent periods less uniformity could be the result of the coexistence of many writing styles.

The most important factors contributing to the separation of literary styles were determined in two distinct ways. The first technique considered a feature to be relevant if it was capable of providing significant distinction between groups, regardless of the other features. The list of metrics and the corresponding p-value for the difference of a given measurement between pairs of clusters are given in Table 3. The asymmetry in the distribution of the average shortest path length  $\varsigma(l)$  and the vocabulary size  $N$  exhibited the most significant variations. Interestingly, similar results were reported in Ref. [12], where these two measurements were also useful to characterize personal writing styles. In the second evaluation, a feature was considered relevant if it was able to provide good distinction between groups based on the interdependencies of features. This evaluation was carried out by computing the importance of each measurement for the axes in the CVA plots. The results in Tables 4 and 5 point to the clustering coefficient ( $C$  and  $C_w$ ) as the main factor for the distinction in 6 clusters. Since there is evidence that the clustering coefficient quantifies whether words are restricted to specific or generic contexts (an explanation of this property is given in Ref. [12])||, it seems that

|| Context-specific restricted words are those appearing in only a few contexts. For example, the concept

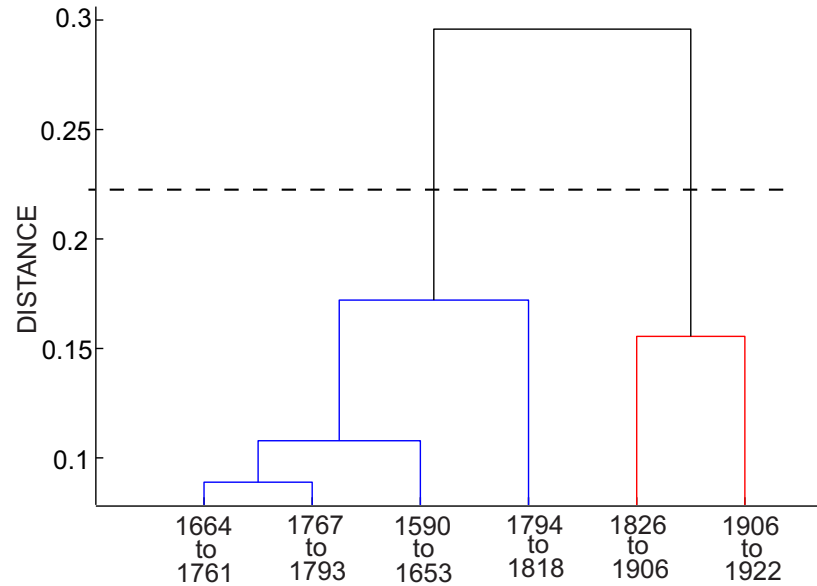


**Figure 5.** Scatter plot representing the best clustering considering the writing style. Note that besides being a good partitioning scheme, it also keeps a good representation of the original database, since 82 % of the variance are kept in the CVA projection.

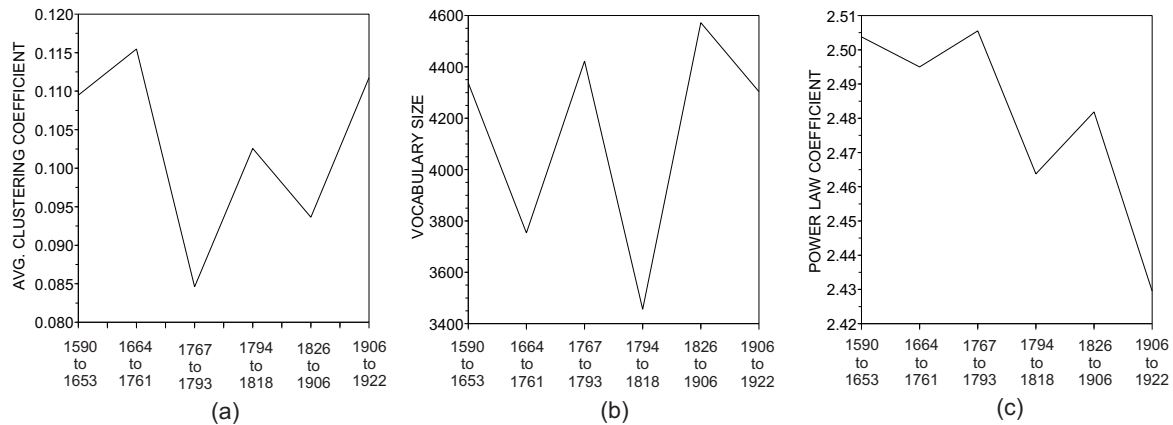
the extent of use of generic or specific words varied along history. This change has not been monotonic, as indicated in Fig. 7(a). In fact, most of the network measurements fluctuated over time, including the size of the vocabulary, whose considerable change was responsible for the most drastic transition, from the 1794 – 1818  $\rightarrow$  1826 – 1906 periods. This is clearly illustrated in Fig. 7(b). The only metric with a well-defined trend over time was the coefficient of the power law for the scale-free networks representing the texts. The decreasing trend in Fig. 7(c) points to a smoother, and therefore more uniform, frequency distribution, which means that the difference in frequency between low and high-frequency words decreased with time.

The changes in style between any two consecutive clusters appeared to have been driven by opposition [40] (see Appendix A), which quantifies the extent into which the current period can be thought of as an opposite movement to the previous literary movements. The coefficient satisfies the inequality  $W_{ij} > 0$ , with the exception of the 1826 – 1906  $\rightarrow$  1909 – 1922 transition. Furthermore, the opposition movement was more significant than the skewness movement  $s_{ij}$  (see Appendix A), which quantifies how much the change in the current style deviates from the opposition movement. The

“teacher” usually induces concepts related to the learning environment. On the other hand, generic words may appear in a myriad of situations. Examples are “red” (red car, red wall or red skin) and “identical” (identical behaviors, identical grades or identical plates



**Figure 6.** Hierarchical relationship between literary periods using the Wards linkage strategy. The 2 groups after the division performed with a particular threshold (dotted line) corresponds to the oldest and to the newest books.



**Figure 7.** Dynamics of (a) average clustering coefficient; (b) vocabulary size; and (c) coefficient of the power law. While the clustering coefficient and the vocabulary size oscillate throughout the periods, the coefficient of the power law tends to decrease, which shows that words were used in a more uniform way in the later periods.

results are given in Table 6. In other words, the innovation of style ( $\vec{v}_i$ , see definition in Appendix A) was generally driven by contrasting the previous styles ( $\vec{a}_i$ , see definition in Appendix A). As for the dialectics  $\rho_{ijk}$  (see Appendix A), which quantifies how the current movement  $i$  is an implication of the two previous movements  $j$  and  $k$ , no clear pattern could be identified in Table 7. The lowest  $\rho_{ijk}$  (and therefore with the highest dialectics) appeared during the 19th century. Thus, realism is a literary style that better

**Table 3.** List of the most significant transitions. Taken individually, the most prominent measurements for discriminating between clusters are the size of the vocabulary  $N$  and the third moment of the average shortest path length  $\varsigma(L)$ .

Measurement	Feature	Transition	p-value
Vocabulary	$N$	1590 – 1653 $\rightarrow$ 1794 – 1818	0.048
	$N$	1664 – 1761 $\rightarrow$ 1767 – 1793	0.051
	$N$	1664 – 1761 $\rightarrow$ 1826 – 1906	0.001
	$N$	1767 – 1793 $\rightarrow$ 1794 – 1818	0.011
	$N$	1794 – 1818 $\rightarrow$ 1826 – 1906	$< 1.0 \cdot 10^{-3}$
Assortativity	$\Gamma$	1590 – 1653 $\rightarrow$ 1767 – 1793	0.008
	$\Gamma$	1590 – 1653 $\rightarrow$ 1826 – 1906	0.044
	$\Gamma$	1664 – 1761 $\rightarrow$ 1767 – 1793	0.041
	$\Gamma$	1664 – 1761 $\rightarrow$ 1826 – 1906	0.006
Shortest Path	$\langle l \rangle$	1664 – 1761 $\rightarrow$ 1826 – 1906	0.049
	$\langle l_w \rangle$	1664 – 1761 $\rightarrow$ 1906 – 1922	0.050
	$\Delta L$	1590 – 1653 $\rightarrow$ 1906 – 1922	0.031
	$\Delta L$	1664 – 1761 $\rightarrow$ 1906 – 1922	0.022
	$\Delta L$	1767 – 1793 $\rightarrow$ 1906 – 1922	0.023
	$\Delta L$	1826 – 1906 $\rightarrow$ 1906 – 1922	$< 1.0 \cdot 10^{-3}$
	$\varsigma(l)$	1590 – 1653 $\rightarrow$ 1826 – 1906	0.028
	$\varsigma(l)$	1590 – 1653 $\rightarrow$ 1906 – 1922	$< 1.0 \cdot 10^{-3}$
	$\varsigma(l)$	1664 – 1761 $\rightarrow$ 1906 – 1922	$< 1.0 \cdot 10^{-3}$
	$\varsigma(l)$	1767 – 1793 $\rightarrow$ 1906 – 1922	0.001
	$\varsigma(l)$	1794 – 1818 $\rightarrow$ 1906 – 1922	0.019
	$\varsigma(l)$	1826 – 1906 $\rightarrow$ 1906 – 1922	$< 1.0 \cdot 10^{-3}$
Clustering	$\langle C \rangle$	1664 – 1761 $\rightarrow$ 1767 – 1793	0.048
	$\langle C \rangle$	1664 – 1761 $\rightarrow$ 1826 – 1906	0.051
	$\langle C_w \rangle$	1664 – 1761 $\rightarrow$ 1767 – 1793	0.054
	$\langle C_w \rangle$	1664 – 1761 $\rightarrow$ 1826 – 1906	0.055
	$\Delta C$	1664 – 1761 $\rightarrow$ 1767 – 1793	0.054
	$\varsigma(C)$	1590 – 1653 $\rightarrow$ 1767 – 1793	0.045

approximates as a synthesis of the two previous literary periods.

In subsidiary studies we verified that the complex network metrics used are indeed efficient in distinguishing styles. For that we examined the writing style dynamics of 10 books¶ of Charles R. Darwin (1809-1882) and Edith Wharton (1862-1937), whose styles are known to differ considerably. Indeed, this is confirmed in the CVA plot in Fig. 8,

¶ The list of books is shown in Table S3 in (SI)-Sec.2.

**Table 4.** Importance of each measurement for the first canonical variable, where the clustering coefficient  $C$  and the average shortest path length  $l$  were the most prominent.

Measurement (First Axis)	Prominence (First Axis)
$\langle C_w \rangle$	33.3 %
$\langle C \rangle$	31.6 %
$\Delta C$	6.6 %
$\langle l \rangle$	6.4 %
$\Gamma$	5.1 %

**Table 5.** Importance of each measurement for the second canonical variable, where the clustering coefficient  $C$  and the average shortest path length  $l$  were the most prominent.

Measurement (Second Axis)	Prominence (Second Axis)
$\langle C \rangle$	34.5 %
$\langle C_w \rangle$	33.7 %
$\langle l_w \rangle$	9.5 %
$\langle l \rangle$	9.4 %
$\Delta C$	3.4 %

**Table 6.** Opposition ( $W_{ij}$ ) and skewness ( $s$ ) indices.

Period	$W_{ij}$	$s_{ij}$
1590 - 1653 $\rightarrow$ 1664 - 1761	1.00	0.00
1664 - 1761 $\rightarrow$ 1767 - 1793	0.39	0.08
1767 - 1793 $\rightarrow$ 1794 - 1818	0.35	0.18
1794 - 1818 $\rightarrow$ 1826 - 1906	1.09	0.07
1826 - 1906 $\rightarrow$ 1909 - 1922	-0.01	0.08

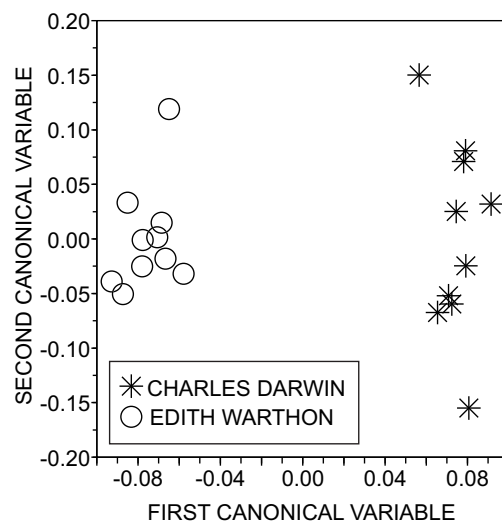
where again the most contributing factor for distinction was the clustering coefficient  $C$ , since both  $\langle C \rangle$  and  $\langle C_w \rangle$  are responsible for 44 % of the weights in the first canonical variable axis.

## 5. Conclusion and further work

Changes in the writing style could be studied objectively by analyzing the metrics from complex networks representing texts from books published over several centuries.

**Table 7.** Counter Dialectics index  $\rho_{ik}$ .

Period	$\rho_{ik}$
1590 – 1653 → 1664 – 1761 → 1767 – 1793	0.76
1664 – 1761 → 1767 – 1793 → 1794 – 1818	1.49
1767 – 1793 → 1794 – 1818 → 1826 – 1906	0.39
1794 – 1818 → 1826 – 1906 → 1909 – 1922	0.69



**Figure 8.** Comparing Darwin’s and Edith Warthon’s styles with CVA projection. A good separation can be observed indicating that these two authors had quite different styles.

Significantly, the most appropriate clustering of books matched the traditional literary classification, with the most contributing factor for distinguishability being the average shortest path length. We found it to be possible to distinguish literary movements using only the vocabulary size or the asymmetry of the average shortest path length distribution. Innovation in writing style was found to be driven mainly by opposition, with growing trend of literary development toward counter-dialectics. Interestingly, these findings represent the generalization of previous results where a dependence was established between network topology and style of machine translations [10, 11] and style of authors [12]. We believe that the approach used here may be useful to study the evolution of any system of interest, since the basic concepts (i.e. characterization through features and use of time series) are completely generic.

As future work, we plan to employ additional complex network measurements in a larger database to verify if the discrimination can be further improved. We shall also

examine the relationship between semantics and topology, by generating clusters using the semantics of words to be compared with the clusters obtained from the analysis of network topology. A more challenging endeavor will be to extend the study to other languages, in order to probe whether the patterns revealed in this paper can be generalized.

**Acknowledgments**

The authors are grateful to FAPESP (2010/00927-9) and CNPq (Brazil) for the financial support.

## Appendix A - Mathematical quantification of writing style

In this appendix we quantify mathematically the variation of writing style. To quantify the change in style over time, we used three concepts, namely *opposition index*, *skewness index* and *counter-dialectics index*, which depend on the measurements computed in each step of the temporal series. For each element  $i$  of the temporal series, which represents the value for the measurements described in Sec. 2.2, we defined the 11-dimensional vector  $\vec{v}_i$ :

$$\vec{v}_i = \left[ N \ \Gamma \ \gamma \ \langle C \rangle \ \langle C_w \rangle \ \Delta C \ \varsigma(C) \ \langle l \rangle \ \langle l_w \rangle \ \Delta l \ \varsigma(l) \right]^T. \quad (6)$$

The large amount of data generated were visualized by projecting  $\vec{v}_i$  into a two dimensional space before computing the indices, and this also helped to remove undesirable correlations. The projection techniques employed are described in (SI)-Sec.3. Using the projected  $\vec{v}_i$ , and considering  $t$  elements in the time series,  $\vec{a}_i$  was defined the average state at time  $i$ ,  $i \leq t$  as:

$$\vec{a}_i = \frac{1}{i} \sum_{j=1}^i \vec{v}_j. \quad (7)$$

Given  $\vec{a}_i$ , the *opposite state* of the current state  $i$  (see Fig. 9(a)) for a geometrical interpretation) is given by:

$$\vec{r}_i = \vec{v}_i + 2(\vec{a}_i - \vec{v}_i) = 2\vec{a}_i - \vec{v}_i, \quad (8)$$

and given  $\vec{r}_i$  and  $\vec{v}_i$ , the *opposition vector*  $\vec{D}_i$  of state  $\vec{v}_i$  (see Fig. 9(a)) is given by:

$$\vec{D}_i = \vec{r}_i - \vec{v}_i. \quad (9)$$

For two consecutive books  $i$  and  $j$ , the vector representing the style change  $\vec{M}_{ij}$  (see Fig. 9(a)) is:

$$\vec{M}_{ij} = \vec{r}_i - \vec{v}_i. \quad (10)$$

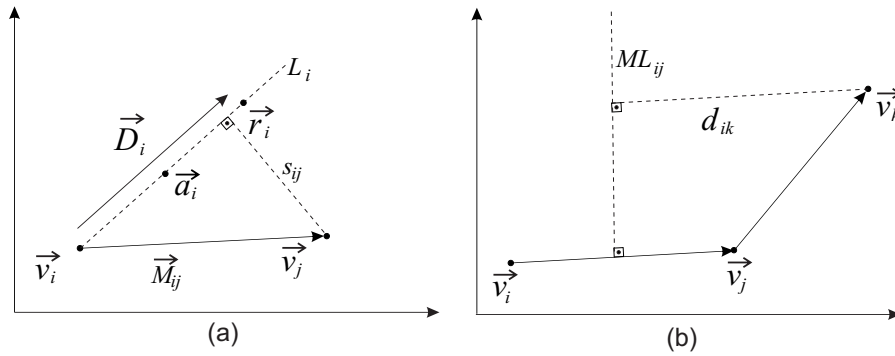
The vector  $\vec{M}_{ij}$  is important because its norm  $\|\vec{M}_{ij}\|$  quantifies the change in style in relation to the previous state  $\vec{v}_i$ . With  $\vec{M}_{ij}$ , the *opposition index*  $W_{ij}$  is the component of  $\vec{M}_{ij}$  over  $\vec{D}_i$ :

$$W_{ij} = \frac{\vec{M}_{ij} \cdot \vec{D}_i}{\|\vec{D}_i\|^2} \quad (11)$$

If the current style tends to oppose the previous one, then the component of  $\vec{M}_{ij}$  over  $\vec{D}_i$  will have a high value. This quantifier is useful, for example, to identify little stylistic innovation: if opposite movements are repeated over and over again, then there is no innovation at all.

The *skewness index*  $s_{ij}$ , which is depicted in Fig. 9(a), is defined as the distance between  $\vec{v}_j$  and the line defined by  $\vec{D}_i$ . This index quantifies how far the stylistic movement is from the opposite movement. It is useful to identify trivial oscillations within the line  $L_i$ , for in this case a series of movements with zero *skewness index* would be observed.





**Figure 9.** Illustration of the quantities employed to define the *opposition*, *skewness* and *counter-dialectics* indices.

The dialectics between three consecutive styles  $i$ ,  $j = i + 1$  and  $k = j + 1 = i + 2$  in the temporal series was quantified as follows. If  $\vec{v}_k$  is the outcome of a synthesis of the styles represented by  $\vec{v}_i$  and  $\vec{v}_j$ , then the distance  $d_{ik}$  between  $\vec{v}_k$  and the middle line  $ML_{ij}$  defined by  $\vec{v}_i$  and  $\vec{v}_j$  (see Fig. 9(a)) will be small. The *counter dialectics index*<sup>+</sup>  $\rho_{ik}$  is:

$$\rho_{ik} = \frac{d_{ik}}{\|\vec{M}_{ij}\|} \quad (12)$$

Further details regarding the definition of the opposition  $W_{ij}$ , skewness  $s_{ij}$  and counter-dialectics  $\rho_{ik}$  are given in Ref. [40].

<sup>+</sup> Note that we referred to  $\rho_{ik}$  as *counter dialectics index* instead of *dialectics index* because it is defined as a distance. Hence, there is an inverse proportion between  $\rho_{ik}$  and the concept of dialectics.

## References

- [1] Michel J B et al. 2011 *Science* **331** 176
- [2] Golder S A and Macy M W 2011 *Science* **333** 1878
- [3] Evans J A and Foster J G 2011 *Science* **331** 721
- [4] Bohannon J 2011 *Science* **330** 1600.
- [5] Newman M E J 2003 *SIAM Review* **45** 167
- [6] Barabási A L 2009 *Science* **325** 412-413.
- [7] Costa L F, Sporns O, Antiqueira L, Nunes M G V and Oliveira Jr. O N 2007 *Applied Physics Letters* **91** 054107
- [8] Ferrer i Cancho R, Solé R V 2003 *Procs. Natl. Acad. Sci. USA* **100** 788
- [9] Antiqueira L, Oliveira Jr. O N, Costa L F and Nunes M G V 2009 *Information Sciences* **179** 584
- [10] Amancio D R, Nunes M G V, Oliveira Jr. O N, Pardo T A S, Antiqueira L and Costa L F 2011 *Physica A* **390** 131
- [11] Amancio D R, Antiqueira L, Pardo T A S, Costa L F, Oliveira Jr. O N and Nunes M G V 2008 *International Journal of Modern Physics C* **19** 583
- [12] Amancio D R, Altmann E G, Oliveira Jr. O N, Costa L F 2011 *New Journal of Physics* (accepted)
- [13] Boginski V L 2005 Dissertation: Optimization and information retrieval techniques for complex networks. University of Florida.
- [14] L Page, S Brin, R Motwani and T Winograd 1999. The PageRank Citation Ranking: Bringing Order to the Web, *Stanford InfoLab*, Technical Report.
- [15] Ratnaparki A 1996 *Proceedings of the Empirical Methods in Natural Language Processing Conference*
- [16] Manning C D and Schütze H 1999 *Foundations of statistical natural language processing* The MIT Press, Cambridge
- [17] Zipf G K 1949 *Human Behavior and the Principle of Least Effort* Addison-Wesley
- [18] Bauke H 2007 *European Physical Journal B* **58** 167
- [19] Newman M E J 2002 *Phys. Rev. Lett.* **89** 208701 s
- [20] Newman M E J 2003 *Phys. Rev. E* **67** 026126
- [21] Newman M E J 2006 *Phys. Rev. E* **74** 036104
- [22] Newman M E J 2010 *Networks: An Introduction* Oxford University Press
- [23] <http://www.gutenberg.org/>
- [24] Ferrer i Cancho R and Solé R V 2001 *Proceedings of the Royal Society of London B* **268** 2261
- [25] Solé R V, Corominas-Murtra B, Valverde S and Steels L 2010 *Complexity* **15** 20
- [26] Stevanak J T, Larue D M and Lincoln D C 2010 *arXiv*: 1007.3254
- [27] Ferrer i Cancho R, Solé R V and Köhler R 2004 *Physical Review E* **69** 051915
- [28] Antiqueira L, Nunes M G V, Oliveira Jr O N and Costa L F 2007 *Physica A* **373** 811
- [29] Roxas R M and Tapang G 2010 *International Journal of Modern Physics C* **21** 503
- [30] Masucci A P and Rodgers G J 2006 *Physical Review E* **74** 026102
- [31] Lee J and Wong D W S 2000 *Statistical Analysis with ArcView GIS* Wiley
- [32] Ward J H 1963 *Journal of the American Statistical Association* **58** 236
- [33] [http://en.wikipedia.org/wiki/Elizabethan\\_era](http://en.wikipedia.org/wiki/Elizabethan_era)
- [34] <http://sparkcharts.sparknotes.com/lit/literaryterms/section5.php>
- [35] <http://en.wikipedia.org/wiki/Neoclassicism>
- [36] [http://en.wikipedia.org/wiki/Age\\_of\\_Enlightenment](http://en.wikipedia.org/wiki/Age_of_Enlightenment)
- [37] [http://en.wikipedia.org/wiki/Gothic\\_fiction](http://en.wikipedia.org/wiki/Gothic_fiction)
- [38] [http://en.wikipedia.org/wiki/Naturalism\\_%28literature%29](http://en.wikipedia.org/wiki/Naturalism_%28literature%29)
- [39] <http://en.wikipedia.org/wiki/Modernism>
- [40] Fabbri R, Oliveira Jr. O N, Costa L F 2010 *arXiv*: 1010.1880